ChinaXiv合作期刊
图书情报工作
第63卷 第22期 2019年11月

新时代人民日报分词语料库构建、性能及应用(一)*
——语料库构建及测评

黄水清[1,2] 王东波[1,2]
[1] 南京农业大学信息科学技术学院 南京 210095 [2] 南京农业大学领域知识关联研究中心 南京 210095

摘要：[目的/意义]构建与新时代相适应的人民日报分词语料库,为中文信息处理提供最新的精标注语料,也为从历时的角度分析现代汉语提供新的语言资源。[方法/过程]在分析已有汉语分词语料库的基础上,描述所构建新时代人民日报语料库的数据源、标注规范和流程,通过构建分词自动标注模型测评语料库的性能,并与已有语料库进行对比。[结果/结论]新时代人民日报语料库遵循现代汉语语料库基本加工规范,规模大,时间跨度长。选取其中的2018年1月部分,基于条件随机场构建分词模型,与1998年1月人民日报语料进行性能测评与对比,所得到的各项具体测评指标表明,新时代人民日报语料整体性能突出,1998年语料无法替代,当前构建该语料库非常必要。

关键词：新时代 人民日报 自动分词 条件随机场模型 语料库 NEPD
分类号：G255.1
DOI:10.13266/j.issn.0252-3116.2019.22.001

1 引言

语料库是由人工或机器标注好的真实语言材料组成的数据集。开展与自然语言相关的研究,语料库是有效的工具和手段。依据语料库既可以研究语言普遍规律也可以针对具体文本开展研究。汉语比其他语种的自然语言处理多自动分词环节,汉语自动分词是一切中文信息处理的基础,汉语分词质量的好坏直接决定了词性标注、实体抽取、自动句法分析和机器翻译等其他中文信息处理任务的性能。目前中文自动分词的主流技术是机器学习,即通过机器学习模型从精加工的语料中自动学习词汇的分布特征和知识,进而完成对汉语字符串中词汇的自动识别,分词语料库是汉语语料库中最重要的类型之一。虽然在同一语料库上基于不同的机器学习模型可以构建不同的分词模型,但整体性能可能差距并不是太大,反倒是训练语料的标注精准度对分词结果影响较大。在中文信息处理的研究中,训练语料通常由通用语料和领域语料组成。在汉语通用语料方面,由北京大学计算语言研究所构建的1998年人民日报语料最具代表性,影响力也最大。但是,随着时间的推移,1998年所构建的精加工人民日报语料,在词汇的时效性、完备性和覆盖度上均需要进行更新、补充和增加。在这一背景下,笔者以2015至2018年《人民日报》发表的文章为对象,构建了新版的人民日报分词语料。因为新版语料库收录的全部是进入新世纪以后的《人民日报》文章,而且均为2012年以后即中国特色社会主义进入新时代以后的文章,为区别于北京大学的1998年人民日报语料,将该语料命名为新时代人民日报语料(New Era People's Daily Segmented Corpus,简称NEPD、NEPD语料或NEPD语料库)。目前NEPD已涵盖了《人民日报》2015上半年(1-6月)及2016年1月、2017年1月、2018年1月共9个月的语料。为促进语料资源的开放和共享,NEPD的相关语料将对学界公布,供学术研究用,并且后续还将不断补充最新语料。NEPD不仅具有动态的历时跨度,而且具有静态的语义丰富度。笔者将用一组文章分别论述NEPD的基本特征、构建过程、分词性能、最佳分词模型,并基于语料从历时的角度探讨当代汉语文本的句式特征、语体特征。本文是其中的第一篇,介绍新时代人民日报语料的构建过程、相应规范和原则,基于条件随机场构建分词模型测评并对比NEPD与1998年1月人民日报语料的性能。结

作者简介:黄水清(ORCID:0000-0002-1646-9300),教授,博士生导师,E-mail:sqhuang@njau.edu.cn;王东波(ORCID:0000-0002-9894-9550),教授,博士生导师。
收稿日期:2019-10-08 修回日期:2019-10-17 本文起止页码:5-12 本文责任编辑:王传清

果对比表明,NEPD 用于处理近年发表的《人民日报》文章时性能明显优于 1998 年人民日报语料,构建 NEPD 非常必要。

2 汉语分词语料及分词模型现状分析

通用汉语分词语料中,具代表性、影响力大的首先是北京大学的人民日报分词语料。该语料库目前发布出来的主要是 1998 年 1 月的人民日报语料,由俞士汶先生带领北京大学计算语言研究所的研究人员完成。该语料库的研制过程中还提出了标注规范,并研究了检索方法[1-2]。其次是国家语委现代汉语通用平衡语料库,该语料库的突出特征是平衡性和规模大,不仅具有新闻语料而且涵盖了经济、军事、体育等不同领域的素材[3]。再次是清华汉语书库中的分词语料,该分词语料的突出特征是基于黎锦熙先生的"凡词,依句辨品,离句无品"的语言学理论实现对汉语分词的[4]。最后是宾州汉语树库中的分词语料,该分词语料库的突出特征是按照结构主义语言学的理论完成对汉语分词的[5]。在上述 4 种汉语分词语料中,前两种分词语料规模较大,所使用的分词理念和规范具有较强的一致性,但是,随着时间的推移,语料时效性问题越来越突出。后两种分词语料所采用的语言学理论具有一定的独特性,但规模上相对较小,且同样存在语料时效性较差的问题。

基于上述对已有汉语分词语料库应用现状及性能的分析,笔者选取 2015 - 2018 年之间共 9 个月的《人民日报》构建新的汉语分词语料库,即 NEPD。NEPD 的构建理由、目标及基本思路是:①从时间上看,中国经过 20 年的快速发展,1998 年所构建的人民日报语料库无论是在词汇的丰富性方面还是覆盖度方面均不能反映当下社会的概貌,需要更新和完善;②《人民日报》在国内外具有很大的影响力,《人民日报》的文章是最为规范和标准的现代汉语,且内容与各时期的中央精神保持高度一致,故仍然选取《人民日报》作为语料库构建的数据源;③1998 年人民日报语料在汉语自然语言处理领域影响力巨大,以最新的《人民日报》为数据源构建新语料库,既延续了前人的成果,也便于开展持续性的研究;④《人民日报》将持续出版,今后可以将新文章不断补充到语料库,扩充 NEPD,使得 NEPD 能够与时俱进,形成能满足时效性要求的实用型大规模现代汉语语料库;⑤相较于汉语词汇的界定,语言学界目前对于汉语词性的数量和分类标准没有达成一致的标准和规范,因此 NEPD 只实现汉语分词,不进行词性标注。

常用的汉语分词机器学习模型主要有隐马尔科夫模型、最大熵模型、条件随机场模型(Conditional Random Fields, CRFs)。在这 3 种模型中,由于条件随机场不仅解决了独立性假设和标记偏置的问题而且在模型训练的过程中能够任意添加特征知识,所以该模型成为了汉语分词的主流技术,比较有代表性的研究如 C. Huang[6]对 1997 年 - 2007 年的中文分词进展进行了回顾,指出相较于手工规则的分词方法,统计学习的分词方法在前者难以解决的未登录词问题上取得了较大突破,是当时的最优解,另外,该作者还强调了公开测评数据集的重要性;与将特征函数定义为二值函数的方法不同,李双龙等[7]将特征函数定义为任意实数值函数从而减少了特征的数量、降低特征选择的复杂度,在 1st SIGHAN 测试集上封闭测试的 F 值为 95.2%;沈勤中等[8]从字的构词能力角度出发,在基础特征的基础上加入字的位置概率特征,实验证明该特征的引入使 F1 值提高了 3.5%,达到 94.5%;迟呈英等[9]在 SIGHAN2006 Bakeoff 的 Uppen、Msra 两种语料的封闭测试中准确率分别达到了 95.8% 和 95.9%,同时也指出条件随机场模型对多字符未登录词的切分效果不佳;宋彦等[10]提出将字、词信息融合的中分分词方法,将条件随机场模型和 Bi-gram 语言模型融为一体,并在 Bakeoff3 上进行封闭验证。最终混合模型效果优于单一模型,F 值达到 93.9%;刘泽文等[11]提出 5 - Tag 标记方法,实验首先采用 LCCRF 模型应用于中文短文本,在此基础上利用词典对初步分词结果进行修正,在 Sighan bakeoff 2005 的 4 个语料测试集上平均 F 值超过 95%,他们的实验表明,加入不合适的特征不但会导致标注结果的 F 值下降,时间复杂度和空间复杂度的上升也更为明显;冯雪[12]利用词典信息设计了一种基于统计的模型,将词典特征融入字的序列标注模型和词的柱搜索模型中,在同领域和跨领域中取得较好的性能;王若佳等[13]结合国内权威词典、官方标准和医学补充词库构建了 10 万数量级的医学辞典,对电子病历进行分词,实现了基于条件随机场的实体识别,F 值达到 82% 的效果,并对识别效果进行了分析。

由于条件随机场模型应用于分词这样的线性序列任务性能较好,本文选择条件随机场模型,以所选取的 NEPD 语料为基础构建分词模型。同时,从评测所构建的 NEPD 语料库的性能角度看,应用条件随机场构建的分词模型便于将基于 2018 年 1 月人民日报语料

黄水清，王东波. 新时代人民日报分词语料库构建、性能及应用(一)——语料库构建及测评[J]. 图书情报工作,2019,63(22):5－12.

构建的分词模型与基于 1998 年 1 月人民日报语料构建的模型进行性能对比。

3 语料获取及预处理

NEPD 的原始语料从《人民日报》图文数据全文检索系统下载得到。所谓原始语料，是指未进行任何标注的、从文本中获取的语言符号的字符序列。为保证 NEPD 语料库在词汇上的覆盖度和历时性，NEPD 的原始语料下载了《人民日报》2015 年 1－6 月、2016 年 1 月、2017 年 1 月、2018 年 1 月总共 9 个月的全部文章。所获取的数据源截图样例如图 1 所示：


授权：南京大学（内部）使用　【全文检索】"(与)、空格(或)　全库检索　高级检索　使用帮助　关闭退出

【返回】【上一记录】【下一记录】【打印预览】【下载记录】【原版样式】　【字号：加大 还原 减小】

日期导航

本篇文章为：人民日报2014.12.31第1版 [版名：要闻] 作者：隋笑飞

新年戏曲晚会在京举行

习近平李克强张德江俞正声刘云山王岐山张高丽出席观看

新华社北京12月30日电 （记者隋笑飞）2015年新年戏曲晚会30日晚在国家大剧院举行。党和国家领导人习近平、李克强、张德江、俞正声、刘云山、王岐山、张高丽等，与首都近千名群众一起观看演出，喜迎新年。

夜幕下的国家大剧院流光溢彩、火树银花。戏剧场内笑语盈盈，暖意融融，洋溢着浓郁的节日气氛。19时45分，习近平等党和国家领导人来到晚会现场，同前来观看演出的老艺术家们一一握手，全场响起热烈的掌声。

今年的戏曲晚会，不同剧种异彩纷呈，传统和新编剧目各具特色，展现出中华戏曲艺术的源远流长和博大精深，将广大戏曲工作者贯彻文艺工作座谈会精神，传承和弘扬中华优秀传统文化的最新成果奉献给了观众。

一部百花争艳的《梨园曲》拉开晚会帷幕。河北梆子《大登殿》、徽剧《小宴》、昆曲《单刀会》等经典唱段展示了传统剧目的无穷魅力。闽剧《丹青魂》、越剧《追鱼》、吉剧《包公赔情》等新编历史剧，将传统与时代相结合，焕发出戏曲艺术的勃勃生机。新编京剧《钦差林则徐》和《安国夫人》一文一武两段表演，技艺精湛、韵味十足。汉剧《红色娘子军》、湘剧《李贞还乡》、眉户戏《迟开的玫瑰》、豫剧《焦裕禄》用古老剧种演绎现代题材，展现优秀党员干部、模范人物崇高的精神境界，更是感人至深。新编京剧、晋剧《廉吏于成龙》表达出自古以来人们对为官者清正廉洁的崇敬。"东方睡狮已醒，中国大船扬帆，华夏儿女同心逐梦，破浪向前……"晚会在京剧歌曲《唱响中国梦》的激昂旋律中结束。整场演出高潮迭起，给观众带来了真善美的艺术享受。

在京中共中央政治局委员、中央书记处书记、全国人大常委会部分领导同志、国务委员、全国政协部分领导同志，有关部门负责同志等观看了演出。

图 1　所获取的《人民日报》数据源截图样例

原始语料获取的具体流程如下：①确定所要获取的《人民日报》原始语料的时间段，并组织人力把时间段内出版的《人民日报》的全部文章从《人民日报》图文数据全文检索系统中下载下来；②把所获得的全部原始语料统一以文本文件形式存储，并保持《人民日报》原有的段落和格式，以方便人工对原始语料进行分词加工和标注；③把所有文本文件统一按月组织在一起，形成给定时间段的完整《人民日报》原始语料。

针对所获取的《人民日报》原始语料，还需要对数据做某些预处理：①需要删除其中不属于《人民日报》正文的内容。人工复制的过程中会把一些非《人民日报》正文的内容复制下来，比如"人民日报 2015.01.27 第 6 版 冯华"这样的内容。对于这些内容可以统一用程序去除。②需要统一《人民日报》原始语料的字符编码。数据获取人员在存储所复制的内容过程中可能会使用不同的字符编码，为了便于后续的统一处理和加工，数据预处理过程中统一将全部数据的编码转化为了 UTF－8 的形式。如此，经过上述数据预处理后，得到了待标注的《人民日报》语料文本。具体样例如表 1 所示：

表 1　《人民日报》待标注语料样例

编号	待标注语料样例
1	有知情人告诉本报记者，建立于 1997 年的证券交易所大楼，在 2002 年时发生过恐怖爆炸事件，也许对建筑本身的牢固性产生了长期的影响。
2	该平台将整合全国农业科教系统相关资源，为广大农民和各类现代农业生产经营主体提供精准、及时、全程顾问式的科技信息服务，促进农业科技创新和成果转化、新型职业农民培育，支撑现代农业的发展。
3	外交部发言人华春莹 25 日宣布，应国务院总理李克强邀请，大不列颠及北爱尔兰联合王国首相特雷莎·梅将于 1 月 31 日至 2 月 2 日对中国进行正式访问并举行新一轮中英总理年度会晤。
4	易司卡尔建议，尽管坍塌对整栋大楼没有太大影响，但如果大楼管理方仍要使用余下的内部通道，就需要对建筑结构进行加固处理，应采取添加支柱等措施，并进行全面安检。

4 语料标注及规范

为保证 NEPD 的标注质量，需要对标注人员进行知识、技能和规范方面的培训，以确保标注人员的整体能力：①所有的标注人员必须掌握如下的知识：有关现代汉语词汇的定义、体系和相关语言学理论；分词在整个中文信息处理研究中的价值和意义；分词不一致的定义和标注；歧义的基本定义及组合型歧义与交集型歧义的区别。②所有标注人员必须掌握自行设计程序实现以下功能的能力：词频统计以及基于齐普夫定律的词频分布规律分析；针对中文的最长匹配分词算法；基于规则的汉语词汇歧义消解算法。③所有标注人员必须系统和完整地熟记国家标准《信息处理用现代汉语分词规范》(GB/T 13715－92)，并能基于该规范中的例子举一反三。

对标注人员完成上述知识、技能和规范方面的培训后，便可通过以下 3 个步骤对经过预处理的《人民日报》原始语料进行人工分词标注。对于每一份原始语料的具体标注步骤如下：①第一组标注人员完成对《人民日报》原始语料的词汇切分。词与词之间切分标记用"/"表示。譬如，"坚持依法治国、依法执政、依法行政共同推进，坚持法治国家、法治政府、法治社会一体建设"，经过第一组的标注后，结果变成"坚持/依法/治国/、/依法/执政/、/依法/行政/共同/推进/，/坚持/法治/国家/、/法治/政府/、/法治/社会/一体/建设/"。②第二组标注人员对第一组的标注结果进行核对。第二组人员需重点关注第一组标注人员是否按照规范对标注对象进行了标注。比如，成语有时被分开标注了："向/全党/提出/扎/扎/实/实/把/全会/提出/的/各项/任务/落到/实处/的/总/要求"，按照标注规范"扎扎实实"应该标注为一个词，正确的标注结果应为

"向/全党/提出/扎扎实实/把/全会/提出/的/各项/任务/落到/实处/的/总/要求"。③第三组人员对经第二组标注人员核对过的分词结果再次进行核对,以确保标注结果的精准性。

经过上述 3 个步骤,《人民日报》原始语料实现了分词标注。为了进一步提升标注结果的精准性,在上述 3 轮标注的基础上,还须设计专门的程序对所有的标点符号进行机器校对,因为标注人员在标注过程中注意力集中在汉语词汇上,容易漏掉对标点符号的标注。

经过上述 3 轮标注和标点的核对之后,最后得到的才是标注完成的 NEPD 语料。具体的标注结果样例如表 2 所示:

表 2 NEPD 语料标注结果样例

编号	标注结果语料样例
1	全面/推进/依法/治国/是/一/个/系统/工程/,/是/国家/治理/领域/一/场/广泛/而/深刻/的/革命/,/
2	冰期/输水/技术/成熟/,/严格/调度/可/在/稳定/冰盖/下/正常/输水/
3	利/节水/、/可/承受/、/保/运行/,/沿线/省/市/根据/实际/制订/居民/水价/方案/。
4	"/芝麻官/、/千钧担/。/作为/县委/书记/,/肩负/着/推动/科学/发展/、/为/民/谋利/造福/的/重任/。/"/广东/省/罗定/市/(/县级/市/)/市委/书记/万/木林/说/

NEPD 语料的标注过程中,在涉及以下几种情形的分词标注中采用了特例规范:①人名在分词过程中采用姓和名分开标注的方式。之所以采用姓与名分开标注的规范,一方面是为了便于以后给姓名添加词性,从而方便统计《人民日报》当中所涵盖的姓氏,另一方面也便于比较 NEPD 与 1998 年人民日报语料中词汇的分布和相应的实验结果。②从语义的组合性上和惯用性方面考虑,在分词标注的过程中将成语看作完整的词汇,但对于字数较多的歇后语、惯用语等,则分开标注成多个词。③对于数与计量单位组合的情形,统一作分词处理。比如,在"个性化地掌握每一名持证残疾人的基本状况"这一表述中,"一名"这一数词与量词的组合应作词切分,具体的标注结果为"个性化/地/掌握/每/一/名/持证/残疾人/的/基本/状况/"。

5 NEPD 分词实验及性能测评

J. Laffrtty 等于 2001 年提出了用于标注和切分序列数据的条件概率模型[14],即条件随机场模型。为测评 NEPD 的分词性能,本文从 NEPD 语料中将 2018 年 1 月的语料单独抽出,与北京大学 1998 年 1 月人民日报语料做分词性能对比。分词实验利用自行封装后的条件随机场开源工具包 CRF + + 0.58 版。CRF + + 使用率较高、可用性较强,特别是在应用于文本处理时易用性、准确率、使用稳定性及通用性等方面均表现突出,并且 CRF + + 的可移植性较强,一般被广泛运用在自然语言处理的分词、命名实体识别及抽取、语义分析等方面。

通过自行开发的封装了 CRF + +0.58 的分词训练和测试平台,首先分别针对 1998 年 1 月和 2018 年 1 月的语料构建自动分词模型,对比它们的性能,然后选取基于 1998 年 1 月语料所构建的性能最好的模型去标注 2018 年 1 月的语料,最后将标注结果与人工构建的 2018 年 1 月的语料进行对比,测评分词性能。通过上述过程,一方面可以验证所构建的新时代人民日报语料的整体性能,另一方面也可以证明构建新时代人民日报语料库的必要性。

5.1 分词实验及性能比较的思路

首先,观察所训练和测试的语料,根据语料表现形式等特点从整体上设计标记符号和特征模板。再分别对所选语料进行相应的标记,将其处理成 CRF + + 能够识别的格式。选取特征并对这些特征进行组合,构造成为相应的特征模板。随后,通过 CRF + + 工具对被选作训练集的数据及特征模板进行处理,得出分词模型,然后对已被同样处理为 CRF + + 可识别格式的测试集数据用所得到的分词模型进行分词处理。输出结果示例如表 3 所示:

表 3 CRF 分词后输出结果示例

文本语料	训练学习标记	测试输出标记
个	B	B
性	M	M
化	E	E
地	S	S
掌	B	B
握	E	E
每	S	S
一	S	S
名	S	S
持	B	B
证	E	E

最后是模型的测评及优化,即利用不同的特征模板训练分词模型,再利用所得到的模型基础上完成对测试语料的标注,并通过测评指标对分词性能进行评测。分词性能的测评指标主要由精准率、召回率和调和平均值(F)构成。具体的计算公式如下:

精准率(P) = 标注正确的标记数/标注为该标记的总数 * 100%

召回率（R）＝标注正确的标记数/应标注为该标记的总数＊100%

调和平均值（F）＝（2＊P＊R）/（P＋R）＊100%

基于上述公式，对不同特征进行实验，得到相应的测评结果，观察它们之间的差异，并根据结果进行特征组合，最终得到分词效果最优的特征选择、特征模板以及对应的分词模型。为更加细致和全面地评估分词模型的性能，不仅需要评估所有标记的标注结果，还需评估多字词（由两个或两个以上汉字构成的词汇）中的单一标记标注结果。在本文实验中，主要使用了构成词的字自身这一单一特征，不涉及音、形等其他类型的特征，在后续的研究中可以增加拼音、部首、字的位置等不同的特征进行分词实验。

5.2 模型性能对比

为构建基于条件随机场的分词模型，并对比时间间隔了20年的新旧两份语料的分词性能，首先分别将2018年1月和1998年1月的语料随机分为10等份，再按照1:9的比例分为测试数据集和训练数据集。在特征选择上，为更加相对公平地对比基于两个不同年份的语料训练得到的模型的性能，仅仅使用构成词的字本身的特征，不添加其他任何特征。不同的标记集合的数量会对模型的性能具有一定的影响，根据人民日报语料中词汇字长的分布情况，标记集合的数量选定为4，因为在中文信息处理的序列化标注任务中，汉语词汇以字为衡量单位的整体长度集中在2－3之间，所以把标记数目限定为4。

具体标记的语义如表4所示：

表4 训练和测试标记含义

标记名称	标记含义
B	B表示词的第一个字
M	M表示词中间的字，并且M可依据词的长度进行无限制使用
E	E表示词的最后一个字
S	S表示单字词字（随着汉语词汇的发展，虽然目前双字词或多字词为主，但仍有一定量的单字词，并且绝大部分单字词的使用频率相对较高）

在语料的训练和测试集中，标记置于所有语料的最后一列。利用CRF＋＋处理训练数据集后，可以得到分词模型，再利用所得到的分词模型对测试数据集进行处理，向测试集输出并添加特征标记序列。在测试语料结果中，所输出的标记序列同样也放置在测试数据集的最后一列，并根据标记的构成情况将字组成词，从而实现对1998年1月和2018年1月人民日报语料的分词。为了便于比较2018年1月和1998年1月两份语料的效果，后续所用的实验所使用的标记和特征模板均是相同的。

基于条件随机场模型，从1998年1月人民日报语料和新构建的2018年人民日报语料中选取不同的等份，按照上述流程，构建得到多个分词模型，并评测对应的精确率、召回率、调和平均值，分别得到10个分词模型，它们在测试语料上的整体性能见表5与表6。

表5 1998年1月人民日报语料的整体性能

模型	评测对象	精准率（%）	召回率（%）	调和平均值（%）
模型1	B	97.14	98.28	97.71
	E	97.20	98.34	97.76
	M	94.44	92.13	93.27
	S	97.62	95.67	96.64
	所有标记	97.10	97.10	97.10
模型2	B	97.17	98.40	97.78
	E	97.13	98.36	97.74
	M	94.69	92.63	93.65
	S	97.75	95.50	96.61
	所有标记	97.13	97.13	97.13
模型3	B	97.17	98.36	97.76
	E	97.15	98.34	97.74
	M	94.75	92.65	93.69
	S	97.66	95.53	96.58
	所有标记	97.12	97.12	97.12
模型4	B	97.02	98.23	97.62
	E	96.99	98.20	97.59
	M	94.50	92.20	93.34
	S	97.52	95.46	96.48
	所有标记	96.97	96.97	96.97
模型5	B	97.13	98.35	97.73
	E	97.13	98.35	97.74
	M	94.55	92.34	93.43
	S	97.76	95.57	96.65
	所有标记	97.12	97.12	97.12
模型6	B	97.31	98.37	97.83
	E	97.20	98.26	97.73
	M	94.53	92.60	93.55
	S	97.67	95.78	96.71
	所有标记	97.18	97.18	97.18
模型7	B	97.18	98.35	97.76
	E	97.18	98.35	97.76
	M	94.40	92.27	93.33
	S	97.70	95.60	96.64
	所有标记	97.12	97.12	97.12
模型8	B	97.24	98.36	97.79
	E	97.13	98.24	97.68
	M	94.39	92.70	93.54
	S	97.80	95.74	96.76
	所有标记	97.15	97.15	97.15
模型9	B	97.15	98.37	97.76
	E	97.11	98.33	97.72
	M	94.58	92.48	93.52
	S	97.66	95.46	96.54
	所有标记	97.09	97.09	97.09
模型10	B	97.15	98.26	97.70
	E	97.05	98.16	97.60
	M	94.32	92.41	93.36
	S	97.50	95.53	96.50
	所有标记	97.01	97.01	97.01

从所有标记的调和平均值的结果来看,基于 1998 年 1 月人民日报语料所构建的分词模型最好性能达到了 97.18%,而所训练的 10 个模型的平均调和平均值为 97.10%。在具体的分词标记上,多字词的首字调和平均值最高性能达到了 97.79%,平均调和平均值为 97.74%;多字词的中间字最高调和平均值为 93.69%,平均调和平均值为 93.47%;多字词的尾字调和平均值最高为 97.76%,平均调和平均值为 97.71%。从多字词的 3 个标记的整体性能看,中间字的性能影响了整个多字词的调和平均值,因为中间字的召回率整体性能较差,最低的召回率仅为 93.27%。跨度比较大的多字词导致了这一问题。比如"沙曼·维雅吉",这是一个人名,本来是一个词,但在所构建的模型中被识别成了"沙曼·维雅"和"吉"两个词。单字词的最高调和平均值为 96.76%,平均调和平均值为 96.61%。虽然单字词的整体性不是最为突出的,但其整体性能分布较为均匀,在一定程度上确保了整个分词模型的性能较为突出。

表 6 2018 年 1 月人民日报语料的整体性能

模型	评测对象	精准率(%)	召回率(%)	调和平均值(%)
模型 1	B	98.08	99.09	98.58
	E	97.96	98.97	98.46
	M	95.28	86.43	90.64
	S	97.85	97.70	97.78
	所有标记	97.80	97.80	97.80
模型 2	B	98.04	99.04	98.54
	E	97.91	98.91	98.40
	M	95.06	86.82	90.75
	S	97.76	97.49	97.62
	所有标记	97.73	97.73	97.73
模型 3	B	98.05	99.02	98.53
	E	97.94	98.90	98.41
	M	94.96	87.10	90.86
	S	97.78	97.52	97.65
	所有标记	97.74	97.74	97.74
模型 4	B	98.01	99.04	98.52
	E	97.87	98.91	98.39
	M	94.90	86.39	90.45
	S	97.79	97.53	97.66
	所有标记	97.70	97.70	97.70
模型 5	B	97.96	98.99	98.47
	E	97.85	98.88	98.36
	M	94.99	86.36	90.47
	S	97.75	97.52	97.64
	所有标记	97.67	97.67	97.67

(续表 6)

模型	评测对象	精准率(%)	召回率(%)	调和平均值(%)
模型 6	B	98.00	99.08	98.54
	E	97.88	98.96	98.42
	M	95.23	86.13	90.45
	S	97.78	97.54	97.66
	所有标记	97.73	97.73	97.73
模型 7	B	98.05	99.09	98.57
	E	97.91	98.95	98.43
	M	95.36	86.72	90.84
	S	97.84	97.56	97.70
	所有标记	97.78	97.78	97.78
模型 8	B	97.99	99.07	98.53
	E	97.85	98.93	98.39
	M	95.24	86.22	90.50
	S	97.78	97.55	97.66
	所有标记	97.71	97.71	97.71
模型 9	B	98.10	99.09	98.59
	E	97.96	98.96	98.46
	M	95.22	86.88	90.86
	S	97.85	97.57	97.71
	所有标记	97.80	97.80	97.80
模型 10	B	98.01	99.05	98.53
	E	97.87	98.92	98.39
	M	94.99	86.46	90.53
	S	97.84	97.56	97.70
	所有标记	97.72	97.72	97.72

基于新构建的 2018 年 1 月人民日报语料,在对所有标记进行评测的基础上,最优模型的调和平均值达到了 97.80%,比基于 1998 年 1 月所构建的最优模型高出了 0.62%。所有标记模型的平均调和平均值达到了 97.74%,比 1998 年 1 月所有模型的平均调和平均值高出了 0.63%。在多字词的首字上,最高调和平均值为 98.59%,平均调和平均值为 98.54%,比 1998 年 1 月的首字平均调和平均值高出 0.8%;多字词的中间字的最高调和平均值达到了 90.86%,平均调和平均值为 90.64%,比 1998 年 1 月的中间字平均调和平均值低了 2.83%;多字词的尾字最高调和平均值为 98.46%,平均调和平均值为 98.41%,比 1998 年 1 月的尾字平均调和平均值高出 0.70%。从历时对比上看,条件随机场模型在中间字的识别方面的性能降低是由于词汇长度跨度增大造成的。在单字词的识别性能上,最高的调和平均值达到了 97.78%,平均调和平均值为 97.68%,比 1998 年 1 月的单字词识别性能分别高出了 1.01% 和 1.17%。

为从模型性能的角度说明构建新时代人民日报语料的必要性,可以从基于1998年1月语料所构建的10个模型中选取调和平均值最高的模型依次去标注2018年1月的10个测试语料。得到的分词标注结果如表7所示:

表7　1998年1月最优模型性能验证

模型	评测对象	精准率(%)	召回率(%)	调和平均值(%)
模型1	B	84.39	92.00	88.03
	E	83.78	91.33	87.39
	M	69.65	32.03	43.88
	S	83.31	82.42	82.86
	所有标记	83.21	83.21	83.21
模型2	B	84.47	91.79	87.98
	E	83.96	91.23	87.44
	M	68.95	31.77	43.50
	S	82.96	82.87	82.91
	所有标记	83.17	83.17	83.17
模型3	B	84.55	91.79	88.02
	E	83.94	91.13	87.39
	M	68.96	32.37	44.06
	S	83.18	82.50	82.84
	所有标记	83.26	83.26	83.26
模型4	B	84.38	91.74	87.91
	E	83.76	91.06	87.26
	M	68.64	32.12	43.76
	S	83.15	82.43	82.79
	所有标记	83.11	83.11	83.11
模型5	B	84.24	91.82	87.87
	E	83.76	91.31	87.37
	M	70.28	31.52	43.53
	S	82.72	82.60	82.66
	所有标记	83.03	83.03	83.03
模型6	B	83.89	91.87	87.70
	E	83.40	91.33	87.18
	M	70.43	30.92	42.97
	S	82.81	82.66	82.74
	所有标记	82.81	82.81	82.81
模型7	B	84.46	92.02	88.08
	E	83.94	91.44	87.53
	M	70.71	31.80	43.87
	S	82.97	82.94	82.96
	所有标记	83.25	83.25	83.25
模型8	B	83.75	90.81	87.14
	E	83.83	90.89	87.22
	M	69.64	32.17	44.01
	S	82.22	82.42	82.32
	所有标记	82.71	82.71	82.71
模型9	B	84.07	91.86	87.79
	E	83.47	91.21	87.17
	M	70.45	30.77	42.83
	S	82.81	83.10	82.95
	所有标记	82.90	82.90	82.90
模型10	B	84.43	91.83	87.97
	E	83.84	91.18	87.36
	M	69.32	32.01	43.80
	S	83.14	82.68	82.91
	所有标记	83.18	83.18	83.18

从表7可以看出,基于1998年1月语料构建的最优标注模型在2018年1月语料上所取得的标注结果与表6基于2018年1月语料所构建的分词模型整体性能差距非常大。所有标记的最高调和平均值为83.26%,平均调和平均值仅达到了83.06%,比基于2018年1月语料所构建模型分别低了14.54%、14.68%。性能指标之所以出现这么大的差异,根本原因是基于20年前的语料所训练出来的模型在词汇的覆盖度和新颖性上已经不能完成对当前文本的精准标注。这也说明,1998年人民日报语料已不适合用于处理当前的最新汉语文本,如果要对当前文本进行自动分词,有必要采用NEPD这样的基于当前文本的新语料。另外,在多字词的中间字的识别方面,基于20年前语料所构建的模型的性能更差,最优调和平均值仅为44.06%,而平均调和平均值也仅为43.62%,与基于2018年1月语料所构建的模型相比分别低了46.80%和47.02%。这一指标表明,基于1998年1月语料所构建的模型用于对2018年1月语料的自动分词时不能解决较长词汇的精准分词问题。

上述实验数据从技术指标方面充分证明了构建新时代人民日报语料的必要性。

6　结语

在分析了目前已有汉语通用分词语料的基础上,本文给出了所构建的新时代人民日报分词语料即NEPD的数据来源、清洗过程、标注规范和标注流程,并从NEPD中选取2018年1月的人民日报语料,通过条件随机场从两个维度验证了所构建语料的整体性能,既证明了NEPD性能突出,也说明了构建该语料的必要性。NEPD可以弥补北京大学人民日报语料用于处理当前文本时的不足。NEPD的构建一方面解决了目前该类分词语料陈旧、过时的问题,从历时的角度实现了对已有人民日报语料的延续和有效扩充,另一方面NEPD可以为开发新的高性能的命名实体识别模型、精准语义检索系统和浅层句法分析器提供有力的资源支撑。后续的研究应注重继续扩大分词语料的规模,并进一步提升语料的精度。

参考文献:

[1] 俞士汶, 段慧明, 朱学锋. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报, 2002(5):49－64.

[2] 王洪俊, 施水才, 俞士汶. 人民日报标注语料的索引方法研究[C]// 全国计算语言学联合学术会议. 全国第八届计算语言学联合学术会议(JSCL－2005)论文集. 南京:南京师范大学,

2005:576 – 578.
[3] 国家语言文字工作委员会. 国家语委现代汉语语料库[EB/OL]. [2019 – 06 – 02]. http://www.cncorpus.org/.
[4] 周强. 汉语句法树库标注体系[J]. 中文信息学报, 2004, 18(4):2 – 9.
[5] ANTONY P J, WARRIER N J, SOMAN K P. Penn treebank-based syntactic parsers for South Dravidian languages using a machine learning approach[J]. International journal of computer applications, 2010, 7(8):14 – 21.
[6] HUANG C, ZHAO H. Chinese word segmentation: a decade review[J]. Journal of Chinese information processing, 2007, 21(3): 8 – 19.
[7] 李双龙,刘群,王成耀. 基于条件随机场的汉语分词系统[J]. 微计算机信息,2006(10): 178 – 180.
[8] 沈勤中,周国栋,朱巧明, 等. 基于字位置概率特征的条件随机场中文分词方法[J]. 苏州大学学报:自然科学版, 2008, 24(3):49 – 54.
[9] 迟呈英,于长远,战学刚. 基于条件随机场的中文分词方法[J]. 情报杂志, 2008, 27(5):79 – 81.
[10] 宋彦,蔡东风,张桂平, 等. 一种基于字词联合解码的中文分词方法[J]. 软件学报,2009(9): 2366 – 2375.
[11] 刘泽文,丁冬,李春文. 基于条件随机场的中文短文本分词方法[J]. 清华大学学报(自然科学版),2015,55(8):906 – 910, 915.
[12] 冯雪. 中文分词模型词典融入方法比较[J]. 计算机应用研究, 2019, 36(1):14 – 16.
[13] 王若佳,赵常煜,王继民. 中文电子病历的分词及实体识别研究[J]. 图书情报工作,2019,63(2):34 – 42.
[14] LAFFRTTY J, MCCALLUM A, PEREIRA F. Conditional random-fields: probabilistic models for segmenting and labeling sequencedata[C]//Proceeding of international conference on machinelearning. Williamstown: International Machine Learning Society, 2001: 282 – 289.

作者贡献说明:
黄水清:提出相关概念及整体研究思路,修订完稿;
王东波:数据处理及初稿撰写。

Construction, Performance and Application of New Era *People's Daily* Segmented Corpus (Ⅰ)
——Construction and Evaluation of Corpus

Huang Shuiqing[1,2]　Wang Dongbo[1,2]
[1] College of Information Science and Technology, Nanjing Agricultural University, Nanjing 210095
[2] Research Center for Correlation of Domain Knowledge, Nanjing Agricultural University, Nanjing 210095

Abstract: [**Purpose/significance**] The construction of the segmented corpus of *People's Daily* in line with the new era provides new annotated corpus for Chinese information processing, and also offers new language resources for analyzing modern Chinese from a diachronic perspective. [**Method/process**] The data source, annotation specification and process of the constructed corpus were explained on the basis of analyzing the existing Chinese word segmentation corpus, on the other hand, the corpus performance was evaluated by constructing the automatic word segmentation model by comparing with the existing corpus. [**Result/conclusion**] The New Era *People's Daily* Segmented Corpus (NEPD) with a large scale and a long time span follows the basic processing standards of modern Chinese corpus. The part of January 2018 is selected from NEPD to build a segmentation model based on conditional random field model. The performance of the corpus of *People's Daily* in January 2018 is evaluated and compared with that of the corpus of *People's Daily* in January 1998. The specific evaluation indexes obtained from the corpus show that the overall performance of the corpus of *People's Daily* in the new era is relatively outstanding. The corpus of 1998 could not be replaced, but it is very necessary to construct the NEPD.

Keywords: new era　*People's Daily*　automatic word segmentation　conditional random field model　segmented corpus　NEPD